Measuring Bayesian (and variational Bayesian) robustness

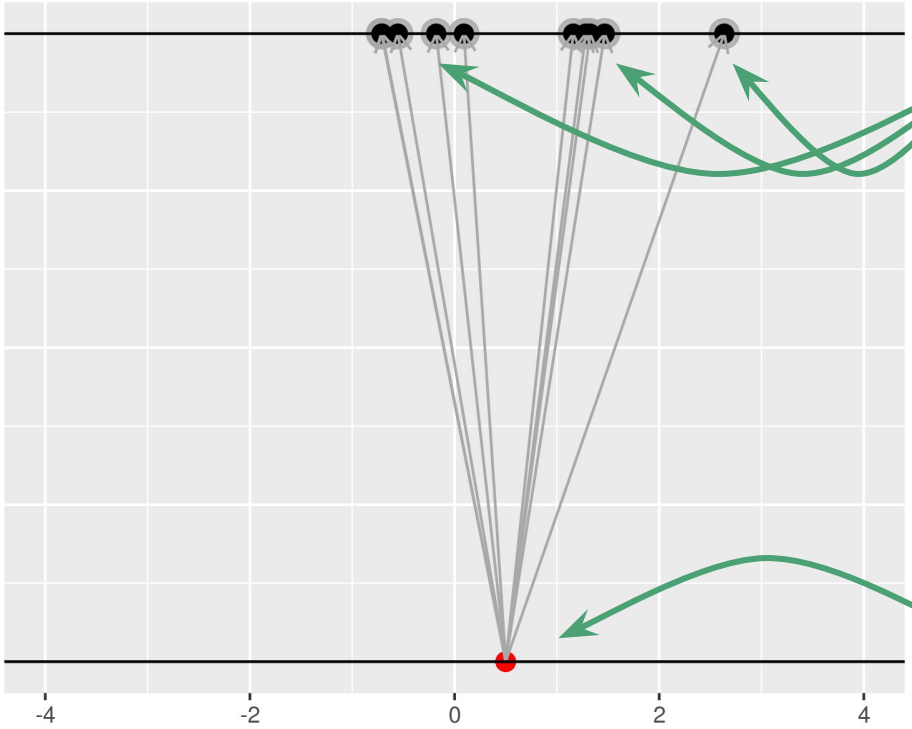
Wow wow

Outline

- What is Bayesian statistics?
  - ...as opposed to frequentist statistics?
  - Bayes rule and posterior approximation
  - Linear regression example
- What do we mean by Bayesian robustness?
  - Local robustness
  - Motivating example: Bayesian “leverage scores”
  - Sensitivity is covariance is sensitivity
  - MCMC for approximate leverage scores?
- What is variational Bayes?
  - The closest distribution in KL divergence
  - Mean field approximations
- How do you calculate robustness to function-valued perturbations?
  - The influence function
  - Worst-case perturbations

What is Bayesian statistics?

Parameters and data



X (data)

Probabilistic model

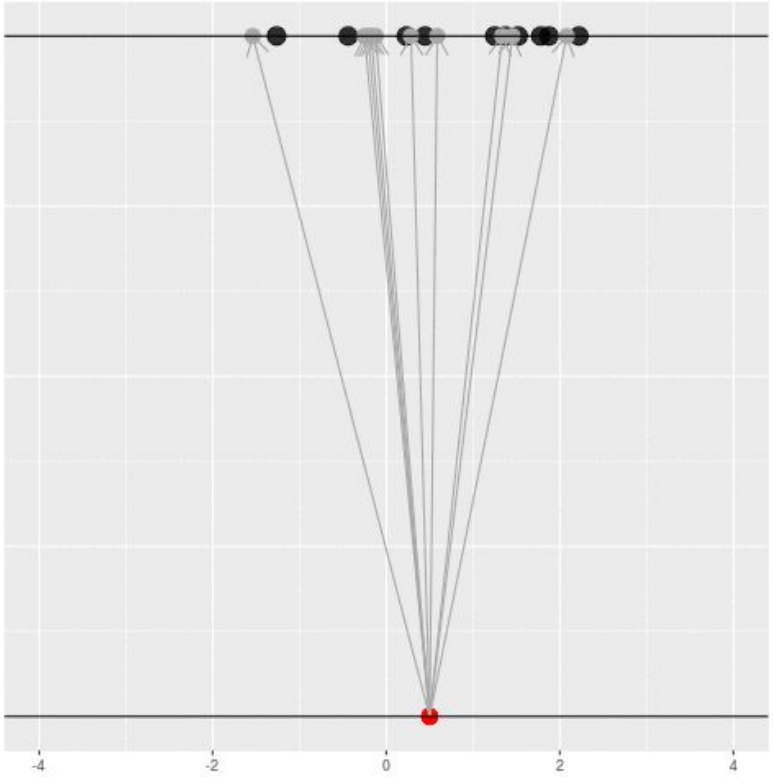
θ (parameter)

Observed

Inference

Unobserved

Frequentist approach



X (data)

Probabilistic model

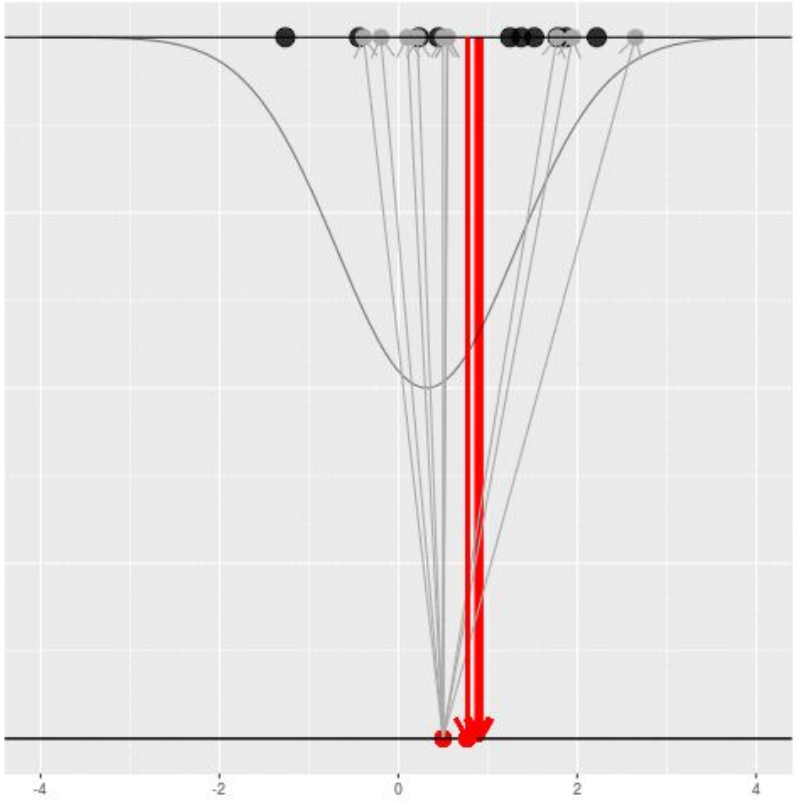
θ (parameter)

Frequentist idea:

We got the parameter indicated by the red dot and saw the dataset in black.

But the same parameter could have given us lots of other datasets.

Frequentist approach



X (data)

Probabilistic model

θ (parameter)

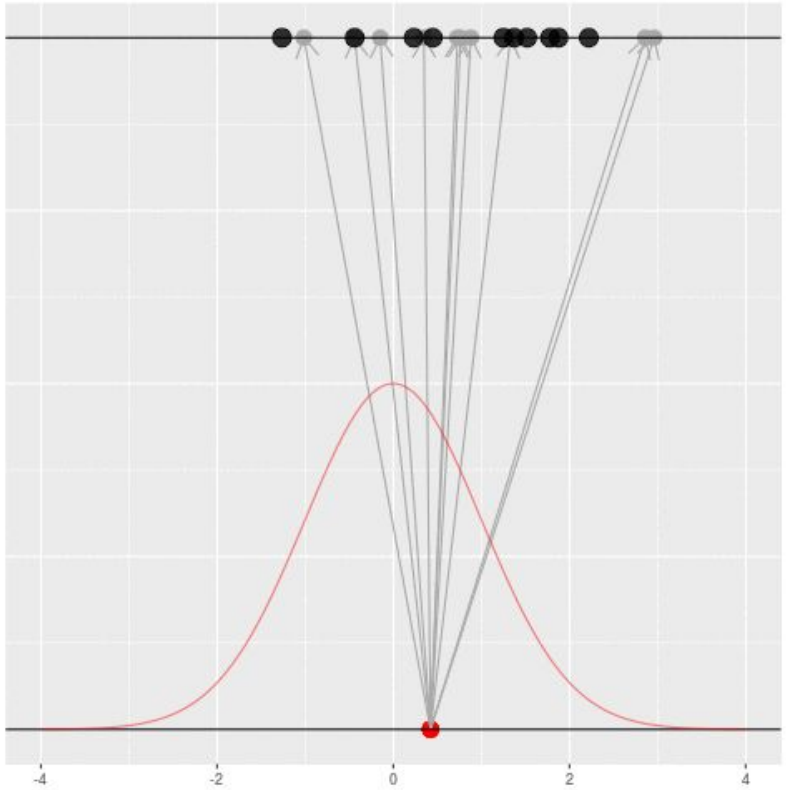
Frequentist idea:

For each dataset, we might pick some summary function and call it an “estimate”.

It will be different each time because the data will be different each time.

We hope the estimate is usually near the true parameter in some sense.

Bayesian approach



X (data)

Probabilistic model

θ (parameter)

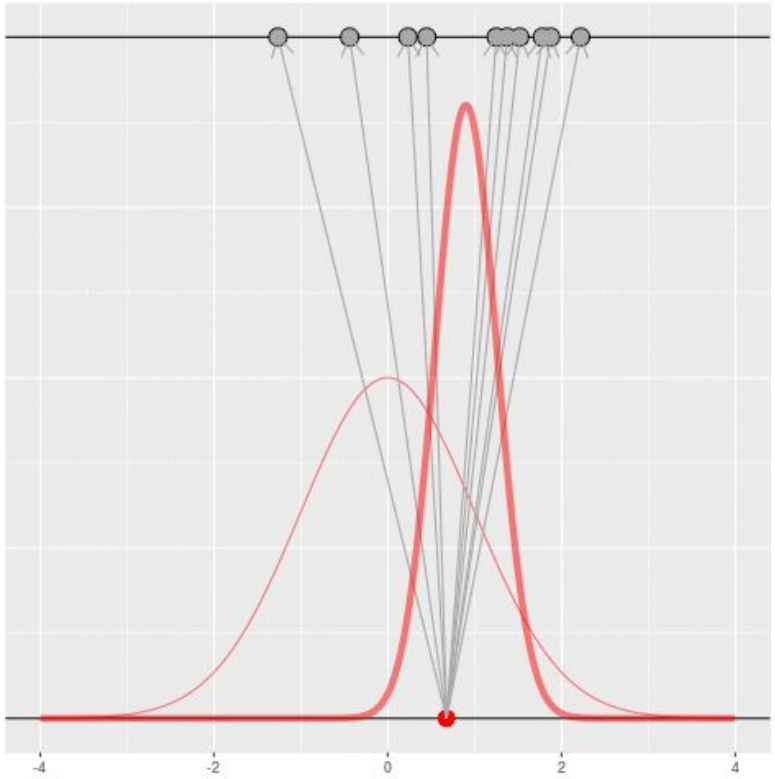
Bayesian idea:

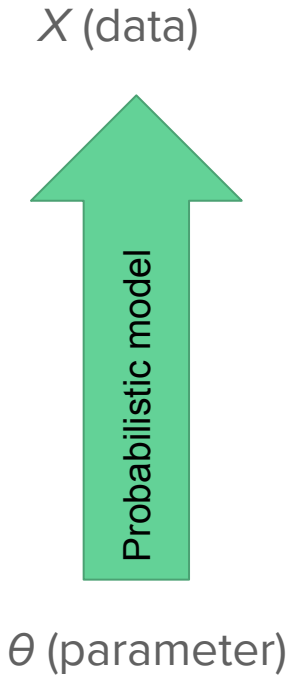
Let's imagine that the universe generates datasets by

- First drawing a parameter from some prior distribution
- and then drawing a dataset from that parameter.

Sometimes we would get the dataset we saw. Mostly we wouldn't.

Bayesian approach





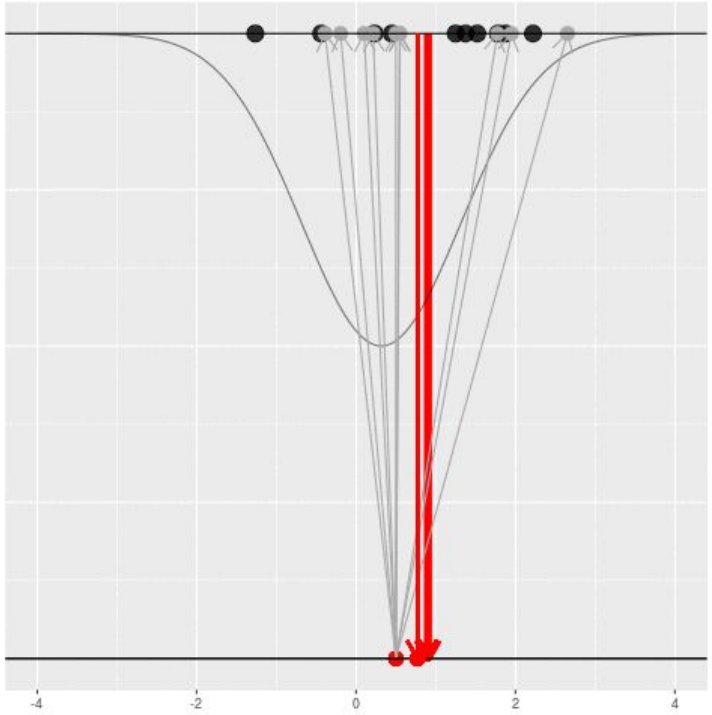
Bayesian idea:

Suppose we draw a bunch of parameters and datasets, and then throw out every pair where the data doesn't match what we observed.

The distribution of the parameters that are left represents which parameters could have given us the dataset we saw.

We hope the prior is reasonable and the model is accurate.

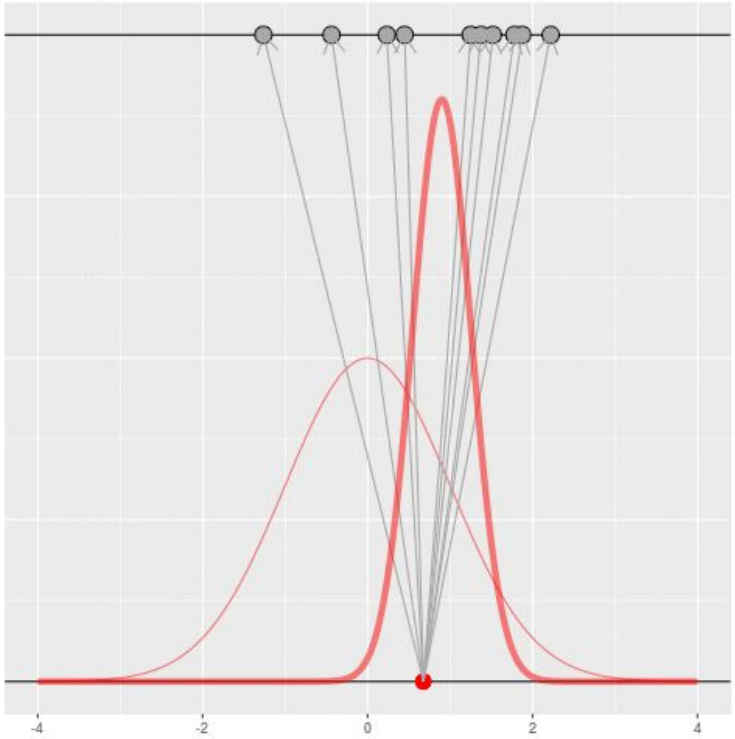
Frequentist: the parameter is fixed, data is random.
Bayes: the data is fixed, the parameter is random.



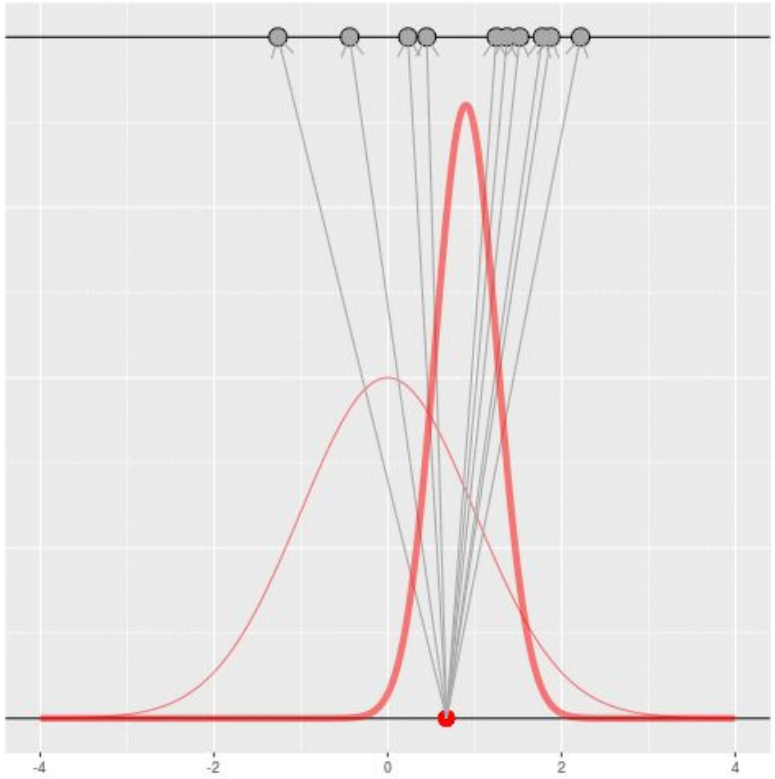
X (data)

Probabilistic model

θ (parameter)



Bayesian approach



Of course, in practice you don't usually generate parameters and data hoping to get your original dataset.

Instead, you use Bayes' rule:

$$p\left(\theta|x\right) = \frac{p\left(x|\theta\right)p\left(\theta\right)}{p\left(x\right)}$$

This is intractable in general (the denominator is a problem). Turn to approximations schemes like MCMC, variational Bayes, &c.

Linear regression example

$$p(y_i|\beta) = \mathcal{N}\left(x_i^T\beta, \sigma^2\right)$$

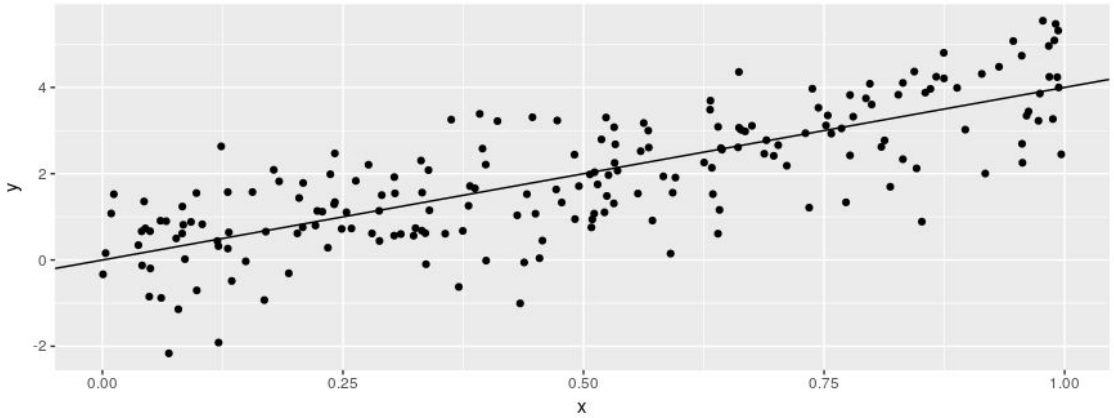
y_i = Observed scalar (data)

β = Unobserved $p-$dimensional column vector (parameter)

x_i = Observed $p-$dimensional column vector

σ^2 = Observed (known) variance

Could think of these as "hyperparameters" -- they are not modeled as random, we assume we just know them



1d example:

We observe the points and want to infer the line.

Linear regression example

We observe *n* data points, so in matrix notation:

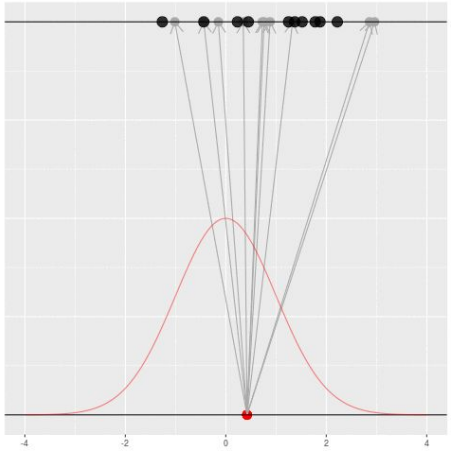
$$p\left(Y|\beta\right) = \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \text{n-dimensional column vector}$$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \text{n by p matrix}$$

Linear regression example

We need to assume a prior.
A convenient (“conjugate”) form is a normal prior:



$$p(Y|\beta) = \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$$
$$p(\beta) = \mathcal{N}\left(0, \sigma_0^2 I_p\right)$$

Then:

$$\log p(Y, \beta) = \log\left(p(Y|\beta)\, p(\beta)\right)$$
$$= -\frac{1}{2}\sigma^{-2}(Y - X\beta)^T(Y - X\beta) - \frac{1}{2}\sigma_0^{-2}\beta^T\beta + C$$

Linear regression example

$$\begin{aligned}
\log p\left(Y,\beta\right) &= \log\left(p\left(Y|\beta\right)p\left(\beta\right)\right) \\
&= -\frac{1}{2}\sigma^{-2}\left(Y-X\beta\right)^{T}\left(Y-X\beta\right)-\frac{1}{2}\sigma_{0}^{-2}\beta^{T}\beta+C \\
&= -\frac{1}{2}\sigma^{-2}\beta^{T}X^{T}X\beta+\sigma^{-2}Y^{T}X\beta-\frac{1}{2}\sigma_{0}^{-2}\beta^{T}\beta+C \\
&= -\frac{1}{2}\mathrm{trace}\left(\left(\sigma^{-2}X^{T}X+\sigma_{0}^{-2}I_{p}\right)\beta\beta^{T}\right)+\sigma^{-2}Y^{T}X\beta+C
\end{aligned}$$

Recall that, for a generic multivariate normal distribution,

$$\begin{aligned}
\log\mathcal{N}\left(\beta;\mu,\Sigma\right) &= -\frac{1}{2}\left(\beta-\mu\right)^{T}\Sigma^{-1}\left(\beta-\mu\right)+C \\
&= -\frac{1}{2}\mathrm{trace}\left(\Sigma^{-1}\beta\beta^{T}\right)+\mu^{T}\Sigma^{-1}\beta+C
\end{aligned}$$

Linear regression posterior

$$\log p(Y, \beta) = \log\left(p(Y|\beta)\, p(\beta)\right)$$
$$= -\frac{1}{2}\text{trace}\left(\left(\sigma^{-2} X^T X + \sigma_0^{-2} I_p\right)\beta\beta^T\right) + \sigma^{-2} Y^T X \beta + C$$

Recall that, for a generic multivariate normal distribution,

$$\log \mathcal{N}(\beta; \mu, \Sigma) = -\frac{1}{2}(\beta - \mu)^T \Sigma^{-1} (\beta - \mu) + C$$
$$= -\frac{1}{2}\text{trace}\left(\Sigma^{-1}\beta\beta^T\right) + \mu^T \Sigma^{-1}\beta + C$$

Therefore the posterior is given by:

$$p(\beta|Y) = \mathcal{N}\left(\left(\sigma^{-2} X^T X + \sigma_0^{-2} I_p\right)^{-1} \sigma^{-2} X^T Y, \left(\sigma^{-2} X^T X + \sigma_0^{-2} I_p\right)^{-1}\right)$$

Linear regression posterior

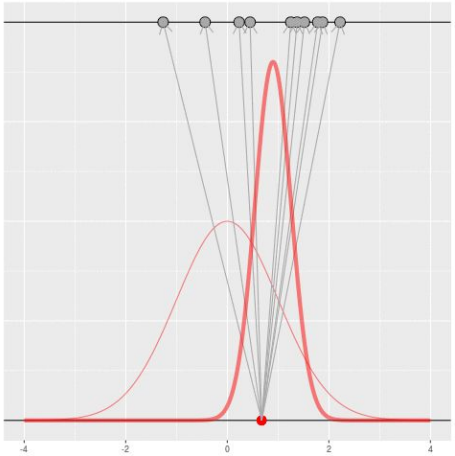
$$p(\beta|Y) = \mathcal{N}\left(\left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\sigma^{-2}X^TY, \left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\right)$$

$$\mathbb{E}[\beta|Y] = \left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\sigma^{-2}X^TY$$

$$\text{Cov}(\beta|Y) = \left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}$$

Taking $\sigma_0 \to \infty$ gives the OLS solution:

$$\begin{aligned}\mathbb{E}[\beta|Y] &= \left(\sigma^{-2}X^TX\right)^{-1}\sigma^{-2}X^TY \\ &= \left(X^TX\right)^{-1}X^TY\end{aligned}$$



Bayesian robustness

Bayesian robustness

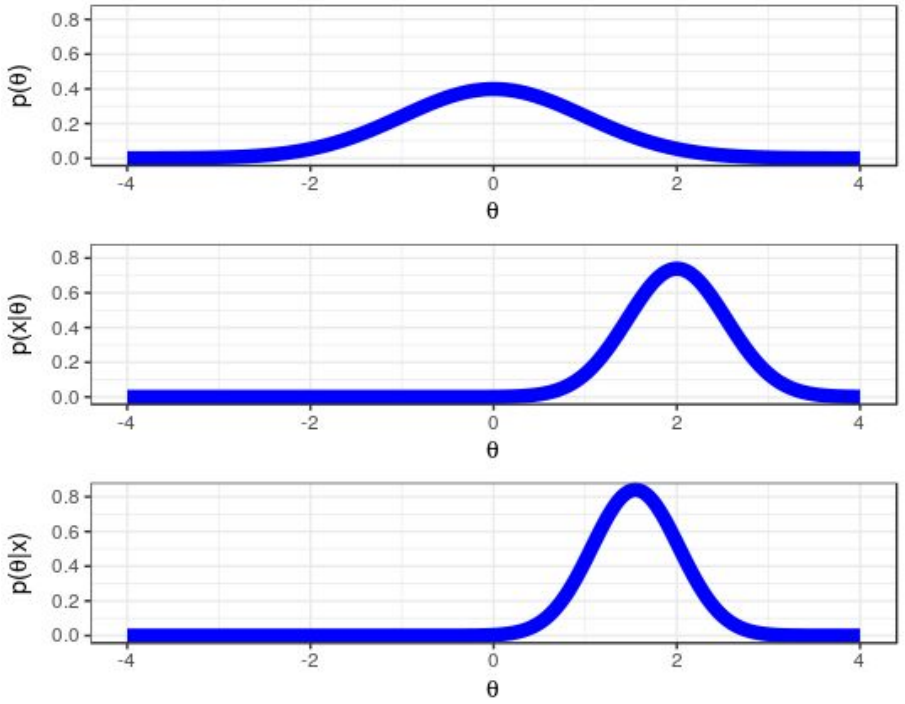
$$p(\theta|\alpha) = \text{Prior}$$

&

$$p(x|\theta) = \text{Likelihood}$$

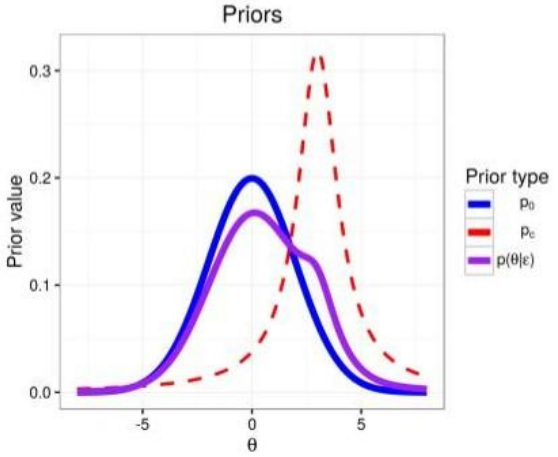
BAYES

$$p(\theta|x,\alpha) = p_{\alpha}^{x}(\theta) = \text{Posterior}$$

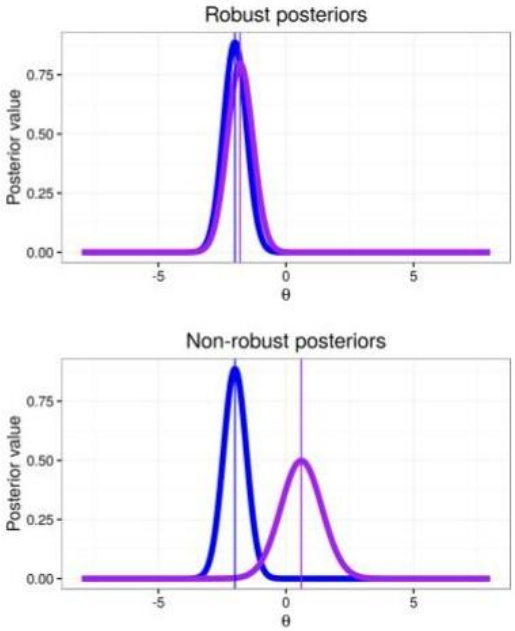


Bayesian Data Analysis
Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin

How do we measure robustness?



Data and Bayes' Rule



Local sensitivity:

Posterior expectation of interest

$$\frac{d\mathbb{E}\left[p_\alpha^x(\theta)\right]}{d\alpha} \quad \text{or} \quad \frac{d\mathbb{E}\left[p_\alpha^x(\theta)\right]}{dx}$$

Perturbation size (of hyperparameters or data)

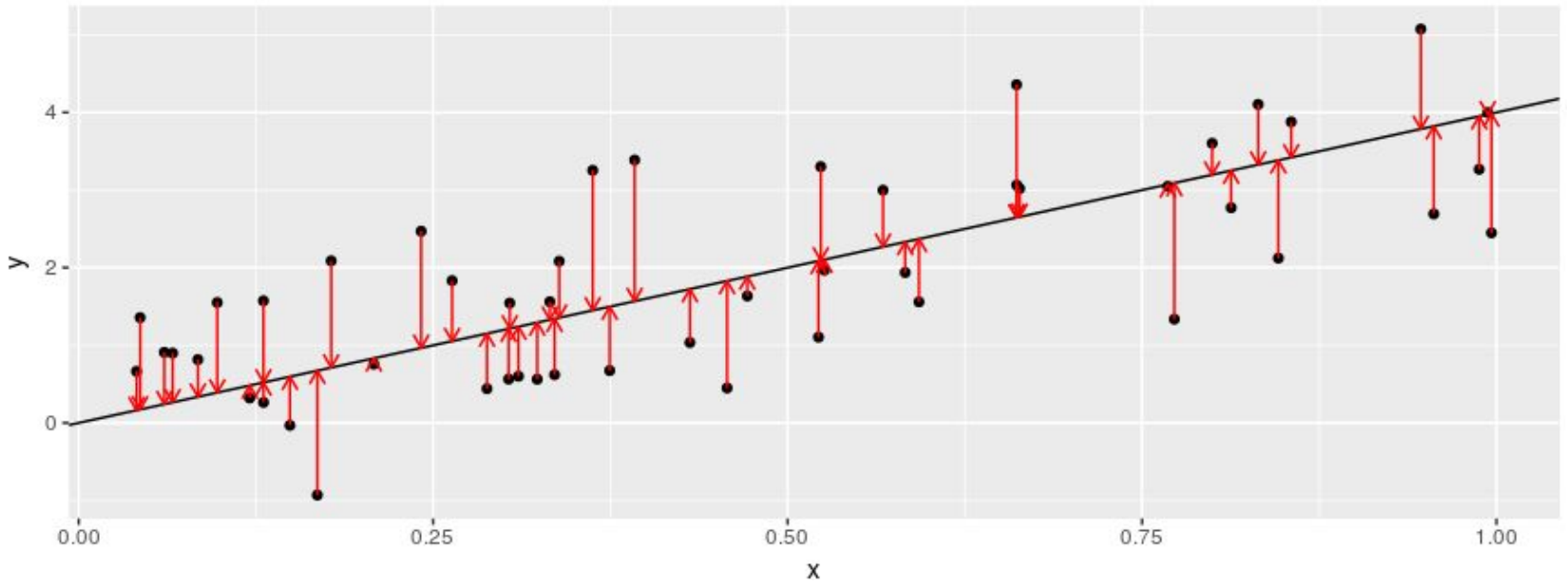
Local Robustness in Bayesian Analysis
Paul Gustafson

Regression example: Bayesian leverage scores

Consider the “fitted values”, or “predictions”:

$$\hat{Y} := X\beta$$

As a function of the unknown parameter, these have a posterior distribution.



Regression example: Bayesian leverage scores

Recall that:

$$\mathbb{E}\left[\beta|Y\right] = \left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\sigma^{-2}X^TY$$

Plugging in:

$$\begin{aligned}\mathbb{E}\left[\hat{Y}|Y\right] &= \mathbb{E}\left[X\beta|Y\right] \\ &= X\mathbb{E}\left[\beta|Y\right] \\ &= X\left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\sigma^{-2}X^TY\end{aligned}$$

Regression example: Bayesian leverage scores

The prediction depends on *Y*, and we can calculate the local sensitivity to *Y*.

$$\mathbb{E}\left[\hat{Y}|Y\right] = X\left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)^{-1}\sigma^{-2}X^TY$$

$$\frac{d}{dY}\mathbb{E}\left[\hat{Y}|Y\right] = X\left(X^TX + \left(\frac{\sigma}{\sigma_0}\right)^2 I_p\right)^{-1}X^T$$

$$\Rightarrow \frac{d}{dy_i}\mathbb{E}\left[\hat{y}_i|y_i\right] = \left(X\left(X^TX + \left(\frac{\sigma}{\sigma_0}\right)^2 I_p\right)^{-1}X^T\right)_{ii}$$

This is the (Bayesian) leverage score!

Sensitivity is covariance is sensitivity

This is a general result, so let's state it in some generality. Choose a family of distributions indexed by *t*:

$$p\left(\theta|t\right) \text{ for some } t \in \mathbb{R}^k$$

We're interested in the sensitivity to *t* of this posterior expectation:

$$\mathbb{E}_{p(\theta|t)}\left[\theta\right] = \int \theta p\left(\theta|t\right) d\theta$$

It turns out that:

$$\frac{d\mathbb{E}_{p(\theta|t)}\left[\theta\right]}{dt} = \mathrm{Cov}_{p(\theta|t)}\left(\theta, \frac{d}{dt}\log p\left(\theta|t\right)\right)$$

Sensitivity is covariance is sensitivity

$$
\begin{aligned}
\frac{d\mathbb{E}_{p(\theta|t)}\left[\theta\right]}{dt} &= \frac{d}{dt}\int \theta p\left(\theta|t\right) d\theta \\
&= \int \theta \frac{d}{dt} p\left(\theta|t\right) d\theta \\
&= \int \theta \frac{d}{dt} \exp\left(\log p\left(\theta|t\right)\right) d\theta \\
&= \int \theta \frac{d}{dt} \left(\log p\left(\theta|t\right)\right) p\left(\theta|t\right) d\theta \\
&= \int \theta \left(\frac{d}{dt}\log p\left(\theta|t\right) - \mathbb{E}_{p(\theta|t)}\left[\frac{d}{dt}\log p\left(\theta|t\right)\right]\right) p\left(\theta|t\right) d\theta \\
&= \int \left(\theta - \mathbb{E}_{p(\theta|t)}\left[\theta\right]\right) \left(\frac{d}{dt}\log p\left(\theta|t\right) - \mathbb{E}_{p(\theta|t)}\left[\frac{d}{dt}\log p\left(\theta|t\right)\right]\right) p\left(\theta|t\right) d\theta \\
&= \mathrm{Cov}_{p(\theta|t)}\left(\theta, \frac{d}{dt}\log p\left(\theta|t\right)\right)
\end{aligned}
$$

(Some regularity conditions are required to differentiate the log probability and to exchange integration and differentiation)

Score is mean zero

Idea: MCMC for approximate leverage scores

Recall the Bayesian leverage score:

$$\frac{d}{dy_i}\mathbb{E}\left[\hat{y}_i|y_i\right] = i^{th} \text{ leverage score} = \mathrm{Cov}_{p_\alpha^x}\left(\hat{y}_i, \frac{\partial \log p\left(Y,\beta\right)}{\partial y_i}\right) = \mathrm{Cov}_{p_\alpha^x}\left(x_i\beta, \sigma^{-2}x_i\beta\right)$$

A step of Hamiltonian Monte Carlo (HMC) requires only the evaluation of:

$$\frac{\partial \log p\left(Y,\beta\right)}{\partial \beta} = -\left(\sigma^{-2}X^TX + \sigma_0^{-2}I_p\right)\beta + \sigma^{-2}Y^TX + C$$

Each requires only O(NP) time.

Idea: MCMC for approximate leverage scores

- Given a matrix *X*, generate data according to *p(Y,β)*
- Run HMC to generate K draws from the posterior *p(Y|β)*
- Calculate the approximate leverage scores using the sample covariance:

$$\bar{\beta} := \frac{1}{K}\sum_{k=1}^{K}\beta_k$$

$$\frac{d}{dy_i}\mathbb{E}\left[\hat{y}_i|y_i\right] = \mathrm{Cov}_{p_\alpha^x}\left(x_i\beta, \sigma^{-2}x_i\beta\right) \approx \frac{1}{K}\sum_{k=1}^{K}\sigma^{-2}\left(\beta_k-\bar{\beta}\right)x_i x_i^T\left(\beta_k-\bar{\beta}\right)$$

The whole procedure should be O(NPK).

What is variational Bayes?

How do you calculate robustness to function-valued perturbations?